国立天文台・天文データセンター　大規模観測データ解析システム II

◯磯貝瑞希、古澤久徳、山根悟、田中伸広、巻内慎一郎、小澤武揚、亀谷和久（ADC）、
大倉悠貴（ハワイ観測所）、髙田唯史（ADC）

概要(Abstract)

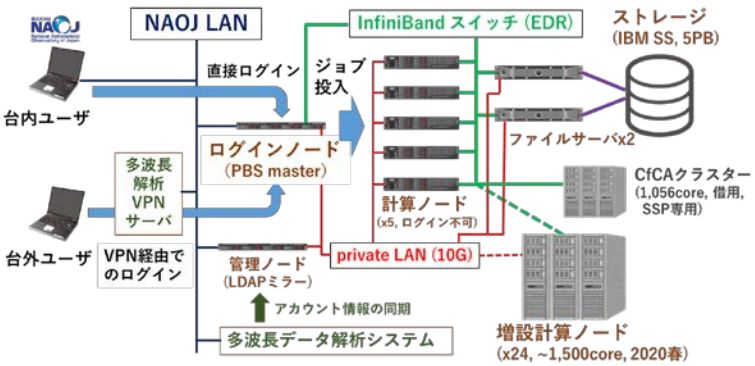
国立天文台・天文データセンター（ADC）では、ハワイ観測所すばる望遠鏡の超広視野カメラ HSC など、解析処理に多くの計算資源を必要とする大規模観測データ用の解析システムを構築し、運用を開始している。本システムは大容量かつ高速 I/O を持つストレージと総コア数 280 の計算ノード他から構成されており、計算ノードの演算性能は今後増強予定である。本講演では、昨年度の講演以降に実施した性能評価試験と初期運用状況について報告する。

1. システムの概要

大規模観測データ解析システム(以下大規模解析)とは、ハワイ観測所すばる望遠鏡の超広視野カメラ HSC など、解析処理に多くの計算資源を必要とする大規模観測データ用の解析システムで、HSC-SSP を含む HSC 共同利用観測者への解析環境提供が初期の主な目的である。このため、初期運用中はシステムを HSC 観測データの解析処理に最適化し、ユーザは HSC 共同利用観測者(PI/CoI)に限定している。本システムの構築は天文データセンター(ADC)、運用は ADC とハワイ観測所(HSC 共同利用+SSP 分)が担当している。

本システムは、ログインノード 1 台、計算ノード 5 台、ファイルサーバ 2 台、ストレージ、管理ノード 1 台で構成される。OS は管理ノードのみ CentOS 7 で、それ以外は Red Hat Enterprise Linux 7 である。計算ノードは 1 台あたり CPU (Intel Xeon gold 6132, 2.6GHz, 14core) を 4 個(合計 56 core/台)、メモリを 1TB 搭載しており、全 5 台の総コア数は 280 である。なお、今後の計算ノード増設でシステムの計算能力を 2,000 コア以上まで増強する予定であり、まずは 2020 年春に 1,500 コア程度の増設を行う。それまで不足する計算能力を補うため、HSC-SSP 用に天文シミュレーションプロジェクトより 1,056 コアのクラスターを借用している。ストレージは容量 5PB でファイルシステムは IBM 社の Spectrum Scale である。図 1 にシステム構成図を示す。

図 1 システム構成図



本システムは、ADC が運用し国内外の研究者へ共同利用サービスとして解析環境を提供している「多波長データ解析システム(以下多波長解析)」と連携しており、システムの利用には多波長解析のアカウントを必要とする。また、計算資源の効率的な利用のため、本システムでは

計算ノードの対話的使用を禁止し、計算資源はジョブ管理ソフトで管理している。ユーザはログインノードからジョブを投入することで計算ノードを使用する。ユーザが利用可能な計算資源とその割り当ての優先度・期間はタイプによって異なる(表 1)。HSC 共同利用観測者は優先度が中（または高）で、100 コア程度の資源を利用可能である。ユーザタイプ毎にジョブ投入に使用するキューを用意しており、現在のキュー構成は表 2 の通りで、HSC 観測者は 1 ジョブ当たり最大で 112 コア、1,780GB のメモリを 15 日間利用可能である。

表 1 利用可能な計算資源・利用期間

利用可能な計算資源とその割当の優先度はユーザタイプによって異なる

ユーザタイプ	利用可能な計算資源	資源割当の優先順位	利用可能なキュー	利用可能期間
HSC-SSP	~ 1,000コア	高	qssp	~2か月x2/年
HSC共同利用観測者（インテンシブ）	~ 100コア	中 - 高	qm /qh	利用宣言後1年間（プログラム終了+1年）
一般（アーカイブ利用者等）	~ 30コア	低	ql	最大1年間（更新可）

表 2 システムのキュー構成（現状）

HSC共同利用観測者 → 利用可能キュー: qm

ユーザタイプ	キュー名	ジョブの優先順位	利用可能なCPUコア数（※1）		利用可能なメモリ量[GB]（※1）		実行可能ジョブ数（※2）		ジョブの最大実行時間
			最大	デフォルト	最大	デフォルト	Hard	Soft	
HSC-SSP	qssp	最高	280	56	4,450	890	---	---	1,000日
HSC共同利用観測者	qh	高	280	56	4,450	890	---	---	1,000日
	qm	中	112	56	1,780	890	---	1	15日
一般（アーカイブ利用者等）	ql	低	28		445		---	1	15日
テストキュー	qt	最高	4		64		1	1	10分

※1 割当計算資源: 増設計算ノードのスペックに応じて変更。
※2 実行可能ジョブ数: 利用者数に応じて調整。

2. 性能評価試験

性能評価およびシステム構築後の動作確認を目的に、3 種類の試験を実施し、先行して納入された計算ノード 2 台で構築したシステム（総コア数:112）および多波長解析（開発系）での結果と比較した。多波長解析（開発系）は CPU (Intel Xeon E5-2667v4, 3.2GHz 8core)を 2 個、メモリを 256GB 搭載したサーバ 3 台から構成されており、以下の(1)と(2)の試験ではその 3 台を、(3)の試験では運用系のバッチサーバの台数と揃えるために 2 台を使用した。試験に用いたストレージは 1 台のサーバと 16Gbps のファイバーチャネル 2 本で接続されており、容量 12TB の領域である。本領域は他のサーバに対しては 10Gbps の LAN を通して NFS ver.3 プロトコルで共有されている。なお、計算ノード 2 ノードと多波長解析（開発系）の結果は昨年度（第 38 回）の技術シンポジウムで報告済みである。

(1) ファイルの Write/Read 速度

ベンチマークソフト IOR を使用してファイルシステムの Write/Read 速度を測定した。表 3 がその結果で、単位は MiB/s、平均値と標準偏差（括弧内に）を掲載している。I/O の並列数は大規模解析 5 ノードで 280、2 ノードで 112、多波長解析（開発系）で 96 で、1 ファイルのサイズは 100MiB、1 I/O Call 当たりのデータ転送サイズは 1MiB、試行回数は 10 回である。大規模解析 5 ノードの W/R 速度は 22-24GiB/s と並列数の増加（112→280）による性能劣化が見られず、多波長解析(開発系)の 10 倍超の性能を維持していることを確認した。

表 3 Write/Read 速度の測定結果

操作	大規模解析 (5ノード, 280core, pIO:280)	大規模解析 (2ノード, 112core, pIO:112)	多波長解析 (3ノード, 48core, pIO:96)
Write	22,765 (101)	23,774 (449)	274 (16)
Read	24,523 (771)	22,449 (845)	1,542 (55)

(2) ファイルのメタデータ操作速度

ベンチマークソフト mdtest を使用してファイルシステムのファイルメタデータ操作（作成、状態確認、読み込み、削除）速度を測定した。表 4 がその結果で、単位は操作数/s、平均値と標準偏差（括弧内に）を掲載している。操作の並列プロセス数は試験(1)と同じで、1 プロセス当たりのファイル作成数は 1,000、試行回数は 3 回である。ファイルメタデータの操作速度は操作内容によって大きく異なるが、全ての操作で並列数の増加（112→280）による性能劣化が見られず、多波長解析（開発系）の 20 倍を超えていることを確認した。以上、試験(1)と(2)より本システムのストレージは少なくとも 280 並列まで高速な I/O 性能を維持することを確認できた。

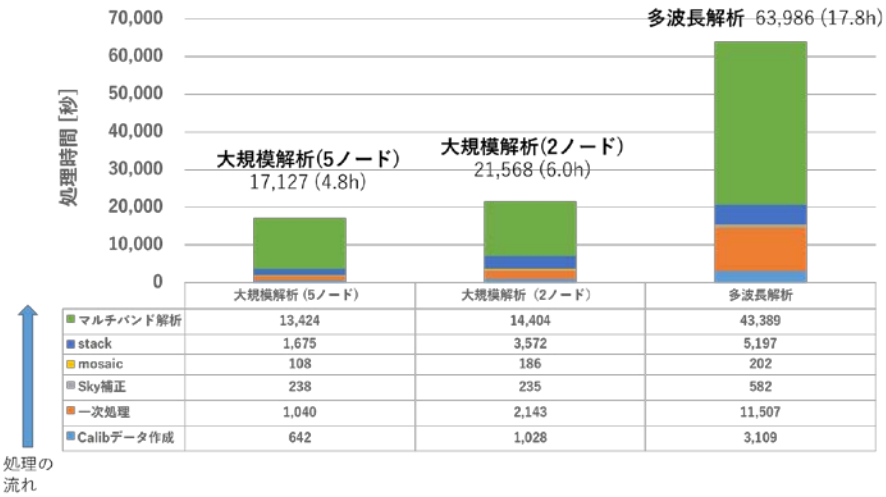
表 4 ファイルメタデータ操作速度の測定結果

操作	大規模解析 (5ノード, 280core, 280p)		大規模解析 (2ノード, 112core, 112p)		多波長解析 (3ノード, 48core, 96p)	
作成	74,345	(25,921)	44,343	(450)	2,106	(60)
状態確認	2,093,134	(27,018)	1,033,556	(178,465)	8,723	(2,499)
読込	282,180	(30,546)	208,597	(6,421)	4,160	(63)
削除	97,213	(56,524)	86,015	(2,131)	3,874	(32)

(3)　HSC データの解析時間

すばる望遠鏡の超広視野カメラ HSC の観測生データからマルチバンド解析による検出天体カタログ作成までの一連の解析処理を行い、その処理時間を測定した。図 2 がその結果である。解析は HSC データの解析処理ソフト hscPipe (ver. 6.7)を使用し、データは HSC 解析講習会のウェブページで公開されているもの(112CCD 版、3 バンド)を用いた。本データの Bias, Dark は 1CCD 当たり 5 枚, Flat, Object は 1 バンド・1CCD 当たり 5 枚(= 3 バンドで 15 枚/CCD)である。処理によっては CCD 数よりも多くのコアを必要としないものがあるため、大規模解析同士での総積算時間の比較では、コア数の差ほどの速度差が見られないが、それでも大規模解析 5 ノードでの解析は、多波長解析のバッチサーバを使用した場合の約 4 倍の速さであることを確認した。

図 2　HSC データ解析時間 [s]



	大規模解析 (5ノード)	大規模解析 (2ノード)	多波長解析
マルチバンド解析	13,424	14,404	43,389
stack	1,675	3,572	5,197
mosaic	108	186	202
Sky補正	238	235	582
一次処理	1,040	2,143	11,507
Calibデータ作成	642	1,028	3,109

3. これまでの運用状況と今後

2019 年 9 月より HSC 観測者(S19B 採択者)の受け入れを開始しており、現在のアカウント保持者:は 7 名である。9 月から 12 月までの 4 か月間で、297 ジョブが実行され、ジョブ実行の積算 CPU 時間は 2,737 日、12 月末時点でのストレージ使用量は 1.3PB と開始早々まずまずの利用状況であった。

今後 2 月より 5 月頃まで HSC-SSP の利用期間に入る。また 2020 年春に総数 1,500 コア程度の計算ノード群を増設する。この増設でシステムの計算能力は現行の約 6.5 倍になる。増設完了後、利用状況を見ながら受け入れ対象を過去の観測者やアーカイブデータ利用者などへ広げることを検討している。